An evaluation of solutions to the problem of boundary change when analyzing long-term relationships on aggregate data

Alvheim, Atle; Olaussen, Thore G.; Sande, Terje

AN EVALUATION OF SOLUTIONS TO THE PROBLEM OF BOUNDARY CHANGE WHEN ANALYZING LONG-TERM RELATIONSHIPS ON AGGREGATE DATA

Atle Alvheim, Thore G. Olaussen, Terje Sande(+)

Abstract: In Norway, communes are the smallest regional political and administrative units, and have existed as such for some 150 years. For this reason the communes have been the main data-carrying unit in the official statistics of the country. This has resulted in a long tradition of well aggregated information at this level.
The Norwegian Social Science Data services has built a database containing part of this information, to further the analysis of regional data. Data may be retrieved for statistical analysis and/or cartographic presentation.
The present article discusses one of the main problems of such a system, changes in regional units over time, and the problems created for analysis of long-term relationships on aggregated data. When changes occur in the system of regional units, the database-system recalculates the data values to the new units. These recalculations are based on information about population transfers involved, and the type of data under consideration. Various underlying assumptions for this procedure are discussed, and so are the different types of error that may introduce into the data. The procedure is tested empirically, and based on the empirical results some recommendations for use are advocated. Since it is possible to recalculate data both forwards and backwards in time, it is recommended that users as a general rule should try to recalculate data following general processes of aggregation in the system of regional units.
Also various types of data do not lend themselves to the same kind of treatment by this procedure. It is mainly designed for variables that give attributes with the population, and is based on the assumption that these attributes approach a homogeneous distribution across the population of a regional unit. The outcome is effected both by the time-period of the retrieved data, i.e. number and types of changes involved, and kind of data retrieved.
The main conclusion is that recalculations of data when units change, to make data comparable, do not seriously affect conclusions based on statistical analyses of all, or large subsets of the regional units. It is more difficult to use long time-series for just a few cases.

GENERAL BACKGROUND

In Norway, communes are the smallest regional administrative units, and they have existed as such for almost 150 years. These units were originally based upon the church subdivision of the country. Accordingly, it is possible to stretch their recorded history even further back in time. As viable political and administrative units the communes have throughout this period been the main datacarrying unit for publication of official statistics, a fact

(+) Address all communications to: Atle Alvheim, Thore G. Olaussen, Terje Sande/NSD, Hans Holmboesgate 22, N-5000 Bergen, Norge

which has given us almost 200 years of good aggregated information at this level.

The Norwegian Social Science Data Services (NSD) has built a database containing part of this information, to further the analysis of communal, and in fact also other types of regional data. To indicate size, the database is now approaching 15000 variables. Figure 1 presents the major parts of this database-system. It is possible to retrieve various types of data as well as documentation. If data is retrieved for further statistical analysis, the user receives a SPSS-file ready for use. There is also an easy link-up with a parallel coordinate-database for the corresponding units and time-periods, making data-display or presentation of analytical results directly available by different types of computer-made thematic maps. This last option has proved itself to be of great value as an extension of the usual statistical methods for analysis of regional data.

In this article we shall discuss one of the main problems of such a system: Handling the problems of frequent boundary changes between regional units when analysing long-term relationships on aggregated data. We shall also present empirical results from tests we have made to evaluate the effects of solutions adopted.

THE SYSTEM

As initially stated, the Commune Data Base (CDB) is a fairly large data-holding, covering the timespan from about 1769 to the present. During this period communes have changed dramatically. In 1769 there were less than 400 communes, increasing over time to some 750 in the 1950's. After a thorough reconstruction of the communes in the 1960's, we are now back to about 450. This creates problems to anyone interested in time-series.

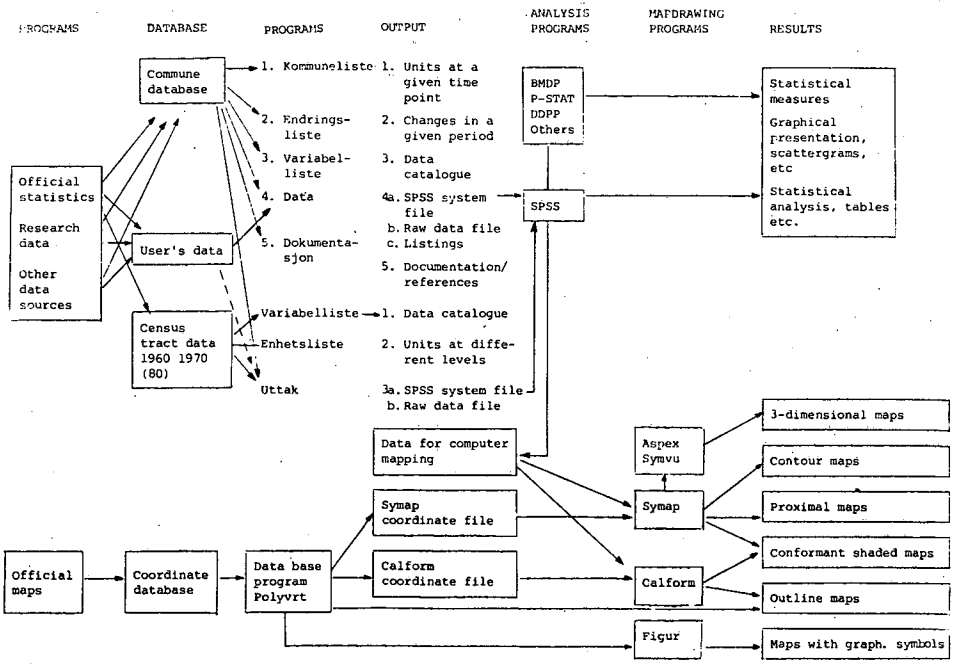
One of the basic principles is that data is always stored on the unit of the year the data is collected. Accordingly, data from 1900 is stored on the 1900-units, while data from 1982 is stored on the 1982-units. Comparison of data from different time-points is therefore often made difficult by the fact that units have changed, so that direct comparison is no longer possible. This is a common problem and we believe that the method we have adopted for solving this problem is of sufficient general interest to justify a somewhat detailed description.

The problem may be solved in at least three different ways:

1. All units involved in changes during the period under study, may be deleted from the dataset. This can be modified by skipping relatively small, insignificant changes. This is at best an unsatisfactory solution to the problem. With data for long timespans or periods with frequent changes in the units, most of the data-matrix will be lost.

2. Another possibility is to aggregate toward larger units. If units are divided, it is possible to aggregate backwards in time, if units are merged we can aggregate forward in time. At least in our Norwegian case, this has proved to be rather complicated. Many changes involve both elements, divisions and mergers. The final result could be relatively large units of litte value to the analyst studying variation: the reduction in number of units would "conceal" much of the information. However, this option is implemented in the latest version of the retrieval-program.

Figure 1: NSD system for regional data



3. The third possibility is to try to measure the amount of resources redistributed when the borders between units are changed. If such measures were available, they could function as standardization coefficients, in principle making it possible to calculate the effect of every change among the units. To be completely correct, one coefficient would be needed for every variable and every change, because changes in the border system between two units do not effect different variables in the same way. Of course, the information necessary to build coefficients at this level of accuracy is simply not available, and if it was, the amount of time needed to construct them would be forbidding. However, one piece of information we have been able to collect and register for every change, is the number of people transferred from one commune to another when borders are changed. Our solution then is to use coefficients based upon the number of people transferred from one commune to another, whenever there is a change in the boundaries. Since most of the variables are attributes with the population, and using the assumption that these attributes approach a homogeneous distribution across the population within each commune, coefficients based upon population transfers will, in most instances, provide us with fairly accurate estimates of redistribution.

Obviously, our assumptions are not always met. In addition, the data needed for the construction of coefficients are in some periods of poorer quality. Especially this goes for the period before 1940, when demographic statistics were not available on a yearly basis.

Standardization of units based on coefficients is intended for use in situations where the explicite goal is to study all or at least a large subset of the communes, in statistical analysis. It is the assumption that errors introduced for a single commune will be leveled out, and not significantly tilt the statistical results. We believe that the effects on statistical measures are smaller in this way than if we just delete troublesome cases from the analysis. Our justification is that different types of communes are treated and effected differently: Cities and towns are expanding, the communes of coastal areas are regrouped according to changes in communication from sea to land, etc.

These coefficients cannot be applied to all types of variables with equal justification. Each variable in the database is therefore associated with a computability-indicator which informs the retrieval-program if and to what extent the coefficients can be used to redistribute values between units.

As we can see, this computability-indicator has to be closely related to the type of data we are going to manipulate. Maybe the easiest to handle is the quite heterogeneous group which we may call global variables. This is data defined for the commune as such. One example may be the communal budget (or most other output from the communal political/administrative process). We have few options: recalculation is usually not possible.

A more difficult group of variables to handle, is what we may term the distributive variables. By this we mean variables for communal level units that are results of aggregations, usually aggregates of individuals in some respect, as inhabitants, an age group, voters or taxpayers, etc. In the CDB, these types of data usually are recalculated when units are involved in changes. However, the picture is made more complicated by the fact that not all the distributive variables are distributions of individuals. It may be farms, houses, factories, etc. as well. For these variables, it is not always possible to use the transfer of people between units as basis for a

recalculation of datavalues. Our assumption that these attributes approach a homogeneous distribution across the population of a commune is certainly violated. So as a first step towards refinement of a crude starting-point, our computability-indicator can take on four values:

Type 0: When units change, variable-values are redistributed on the basis of the coefficients. Usually this is data based on aggregation of individuals.

Type 1: Variables of this sort will only be recalculated when whole units are merged, and then by adding the values. This applies to situations where it is not possible to say that there is a direct relation between transfer of population and transfer of the variable under consideration. For example, surface cannot be recalculated on the basis of population changes, but if two or more communes are totally merged, it is possible to calculate the correct value merely by adding. Otherwise, the variable will get missing data values for the units involved in the change.

Type 2: This is a type of variables where units involved in a change are checked for the same value on the variable under consideration. Examples may be indices or typologies. If two communes being merged have the same value on a typology, the new unit(s) receive(s) this value, otherwise a missing data code is inserted in the data.

Type 3: No recalculation whatsever is possible. This usually goes for variables measuring attributes with the communes as an administrative entity. Examples are budgets or decisions.

When the user of this system wants to retrieve data s/he has to submit a time-point and list of variables to be retrieved. The time-point serves as standardization-year for the unit-dimension of the output-file. That means, if we specify 1980, all the variables of our variable-list are recalculated to the 1980-units (communes). If we want to compare census-data from 1950, 1960, 1970, and 1980, the 1950-data on the 746 1950-communes will be recalculated to the 454 1980-comnmunes, and vice versa.

To sum up: as the starting point, the user in such a situation would have two basic possibilities:

a) S/he may aggregate to the smallest common denominator, or
b) Data may be recalculated according to the rules of the system.

The sophisticated user would choose according to

a) Kind of data retrieved, and
b) The time-period of the retrieved data, i.e. number and types of changes involved.

The computability-indicators associated with the variables make this choice easier for the user.

It is possible to recalculate data both forwards and backwards in time. One of our general recommendations is that it is usually preferable to recalculate towards fewer units. Data become more reliable if we follow a general process of aggregation.

In addition, we must be aware of two different types of errors introduced:

a) A variable to be recalculated may be skewed, - not homogeneously distributed across a commune.

b) The number of people living in the transferred part of a commune may not be a constant percentage of the total population of that commune over time, but increase much faster than in the rest of the communal area. If we try to recalculate data from the 1900-census to 1970-units, and the commune is involved in a change in 1965, we have to recalculate 1900-data with a coefficient based on the 1965-demography. This is a serious problem when we work with long time series, or if change does not mean total merge of units.

For a clearer picture of these effects, we have tried to test this procedure empirically. We have chosen the most difficult time-period, 1960-1970, and tried to answer the following questions:

1) How accurate are the actual data values being recalculated?

2) Are there any differences in using actual values compared to relative values?

3) What happens to the relations between two or more variables:
   - when the variables under considerations are measures of the level of some attribute;
   - when variables are measures of change?

4) These results should be compared to the effects of the alternate method, aggregation to smallest common denominator.

Test of recalculation using comparisons with actual data

This sort of test can easiest be made for the time-period after 1950, when census-data started to be published on census-tract level. Census-tracts usually were the building-blocks used later on when communes were restructed, by adding or subtracting data for census-tracts it is possible to compare data from the 1960- and the 1970-censuses for the same physical units. The very thorough reconstruction of communes from 1960 to 1970 almost completely fulfill this requirement. Because of no troublesome deviations, we have in this test focused our empirical work on the Stavanger area of Rogaland county. The variable we have picked is the number of people employed in industry. This variable is deemed specially sensitive to all the problems of recalculation we already have pointed to. However, as a control, we have also selected another (less sensitive?) variable to be tested in parallel: Number of women 30-49 years of age.

Table 1 shows first the actual and relative figures for the 1960-communes, with the three columns showing actual data 1960, actual data 1970 on the census-tracts that make up the 1960-commune, and the 1970-data recalculated to the 1960-commune using coefficients.
The second part of the table shows the same the other way around, 1960-data on 1970-units.

Lastly, the third part gives the same information when the method of smallest common denominator (SCD) is used. Actually, three different sets of units have been tested. In the first set we ignored all changes effecting less than 10 % of the total population of a commune. This leaves us with four remaining units. In the second set we have fixed this limit at 5 %, which results in three units for further analysis.

Table 1: Actual and recalculated values for communes in the Stavanger - Sandnes area

1960-comm.	Employed in industry			% of all employed			Women 30-49 years			% of tot. population		
	1960	1970		1960	1970		1960	1970		1960	1970	
		Actual	Recalc.		Actual	Recalc.		Actual	Recalc.		Actual	Recalc.
Stavanger	9233	6591	7701	43.6	38.9	39.0	7458	4433	6060	14.2	9.9	11.3
Madla	645	1596	902	38.0	37.9	39.0	693	1631	709	14.7	13.0	11.3
Hetland	3235	3907	3558	41.9	40.1	40.0	2958	3387	2723	14.6	12.6	11.6
Sandnes	859	642	833	51.3	45.9	49.6	541	317	510	13.6	9.3	14.7
Høyland	3951	4603	4283	55.0	50.9	49.6	2640	2899	2620	14.1	16.7	14.7
Høle	101	136	205	29.5	40.1	50.1	105	74	123	11.2	7.8	14.5
Forsand	184	231	288	31.0	45.2	49.7	190	142	147	11.3	9.5	9.3
Gjesdal	834	948	877	62.8	60.3	58.9	399	358	349	12.5	9.4	9.5
Bjerkreim	70	133	140	11.2	18.0	18.8	209	171	171	11.5	9.0	9.0

1970-comm.	Employed in industry			% of all employed			Women 30-49 years			% of tot. population		
	1960		1970	1960		1970	1960		1970	1960		1970
	Actual	Recalc.		Actual	Recalc.		Actual	Recalc.		Actual	Recalc.	
Stavanger	12867	12820	11724	42.6	42.9	39.0	10862	10841	9225	14.3	14.3	11.3
Sandnes	5153	5200	5748	53.7	52.6	49.6	3531	3550	3516	13.9	14.0	14.7
Forsand	109	115	126	28.7	30.8	41.3	120	119	82	11.2	11.3	9.1
Gjesdal	915	909	1059	58.4	57.7	58.9	476	479	422	12.3	12.3	9.5
Bjerkreim	68	68	130	11.1	11.1	17.9	204	204	167	11.5	11.5	9.0

	Smallest common denominator	Employed in industry			% of all employed			Women 30-49 years			% of tot. population		
		1960		1970	1960		1970	1960		1970	1960		1970
		Actual	Aggreg.		Actual	Aggreg.		Actual	Aggreg.		Actual	Aggreg.	
10% limit	Stavanger Madla Hetland	12867	13113	11724	42.6	42.8	39.0	10862	11109	9225	14.3	14.3	11.3
	Sandnes Høle Høyland	5153	4911	5748	53.7	53.4	49.6	3531	3286	3516	13.9	13.9	14.7
	Gjesdal Forsand	1024	1018	1185	52.6	53.0	56.4	596	589	500	12.0	12.1	9.5
	Bjerkreim	68	70	130	11.1	11.2	17.9	204	209	167	11.5	11.5	9.0
5% limit	Stavanger Sandnes Hetland Høyland Høle Madla	18020	18024	17472	45.3	45.3	41.9	14393	14395	12741	14.2	14.2	12.1
	Gjesdal Forsand	1024	1018	1185	52.6	53.0	56.4	596	589	500	12.0	12.1	9.5
	Bjerkreim	68	70	130	11.1	11.2	17.9	204	209	167	11.5	11.5	9.0
No limit	Tot. area	19112	-	18787	45.1	-	42.2	15193	-	13408	14.1	-	11.9

Lastly, if the limit is set to zero, all changes will be taken into account, and we are left with only one unit. To help readers better evaluate the results in the tables, figure 2 presents a schematic picture of the communal rearrangements for this area during the period 1960-1970.

Figure 2: Picture of communal rearrangements in the Stavanger area from 1960 to 1970

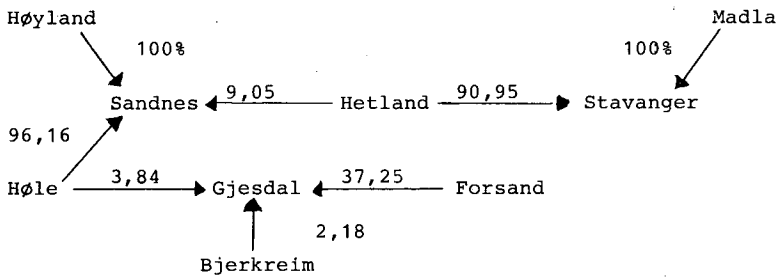


Table 2 is a condensed version of table 1. It presents the absolute difference between the actual and recalculated values, this difference as a percentage of actual value, and thirdly, the difference between the two relative versions of the variables.

If we only look at size, it is obvious that the errors are smaller the bigger and/or fewer units involved in the calculations. But it is as obvious that variance will be lost, big units certainly make for less homogeneous units. Some of the variance between units is turned into variance within units. Second, there seems to be only insignificant differences between the results for the two variables, actually it might look like it is a small difference "the wrong way", the postulated most sensitive variable may in fact have been the least sensitive. Thirdly, it is certainly better to calculate towards fewer units, if we want the data to be as correct as possible. To merge units does not produce much error, and to calculate towards fewer units almost always involve some plain adding. Fourth, we introduce less error if we use recalculated variables as relative measures when possible.

Table 3 presents measures of change from 1960 to 1970 for the same units and variables. The first two columns give yearly rate of change based on absolute values, first the actual (correct) data and next the recalculated data. The next two columns give the difference between 1960 and 1970 measured as percentage points, first the actual data and then the recalculated data.

It is difficult to draw strong conclusions from these few units, but it seems to be a general tendency that:

1. The deviation between the rate of change based on actual and recalculated data is greater than if we look at the level.

2. The change-rates for actual and recalculated date seem to differ more

Table 2: Difference between actual and recalculated values for the Stavanger - Sandnes area

1960-comm.	Employed in industry Disparity 1970-data Absolute	Employed in industry Disparity 1970-data Percentage	% of all employed Disparity 1970-data Percentage-points	Women 30-49 years Disparity 1970-data Absolute	Women 30-49 years Disparity 1970-data Percentage	% of tot. population Disparity 1970-data Percentage-points
Stavanger	1110	16.8	0.1	1627	36.7	1.4
Madla	- 694	-43.5	1.1	- 922	-56.5	- 1.7
Hetland	- 349	- 8.9	- 0.1	- 644	-19.6	- 1.0
Sandnes	191	29.8	3.7	193	60.9	5.4
Høyland	- 320	- 7.0	- 1.3	- 279	- 9.6	- 2.0
Høle	69	50.7	10.0	49	66.2	6.7
Forsand	57	24.7	4.5	5	3.5	- 0.2
Gjesdal	- 71	- 7.5	- 1.4	- 9	- 2.5	0.1
Bjerkreim	7	5.3	0.8	0	0.0	0.0

1970-comm.	Employed in industry Disparity 1960-data Absolute	Employed in industry Disparity 1960-data Percentage	% of all employed Disparity 1960-data Percentage-points	Women 30-49 years Disparity 1960-data Absolute	Women 30-49 years Disparity 1960-data Percentage	% of tot. population Disparity 1960-data Percentage-points
Stavanger	- 47	- 0.4	0.3	- 21	0.0	0.0
Sandnes	47	0.9	- 1.1	19	0.5	0.1
Forsand	6	5.5	2.1	1	- 0.8	0.1
Gjesdal	- 6	- 0.6	- 0.7	3	0.6	0.0
Bjerkreim	0	0.0	0.0	0	0.0	0.0

	Smallest common denominator	Employed in industry Disparity 1960-data Absolute	Employed in industry Disparity 1960-data Percentage	% of all employed Disparity 1960-data Percentage-points	Women 30-49 years Disparity 1960-data Absolute	Women 30-49 years Disparity 1960-data Percentage	% of tot. population Disparity 1960-data Percentage-points
10% limit	Stavanger Madla Hetland	246	1.9	0.2	247	2.3	0.0
	Sandnes Høle Høyland	- 242	- 4.7	- 0.3	- 245	- 6.9	0.0
	Gjesdal Forsand	- 6	- 0.4	0.4	- 7	- 1.2	0.1
	Bjerkreim	2	2.9	0.1	5	2.5	0.0
5% limit	Stavanger Sandnes Hetland Høyland Høle Madla	4	0.0	0.0	2	0.0	0.0
	Gjesdal Forsand	- 6	- 0.4	0.4	- 7	- 1.2	0.1
	Bjerkreim	2	2.9	0.1	5	2.5	0.0
No limit	Tot. area	-	-	-	-	-	-

Table 3: Measures of change based on actual and recalculated values for the Stavanger - Sandnes area

	1960-comm.	Change 1960-70, industrially employment: Yearly rate of change Actual	Yearly rate of change Recalc.	Percentage-points Actual	Percentage-points Recalc.	Change 1960-70, women 30-49 years: Yearly rate of change Actual	Yearly rate of change Recalc.	Percentage-points Actual	Percentage-points Recalc.
	Stavanger	- 3.3	- 1.2	- 4.7	- 4.6	- 5.1	- 2.0	- 4.3	- 2.9
	Madla	9.5	3.4	- 0.1	1.0	8.9	0.2	- 1.7	- 3.4
	Hetland	1.9	1.0	- 1.8	- 1.9	1.4	- 0.8	- 2.0	- 3.0
	Sandnes	- 2.9	0.0	- 5.4	- 1.7	- 5.2	- 0.6	- 4.3	1.1
	Høyland	1.5	0.8	- 4.1	- 5.4	0.9	0.0	2.6	0.6
	Høle	3.0	7.3	10.6	20.6	- 3.4	1.6	- 3.4	3.3
	Forsand	2.3	4.6	14.2	18.7	- 2.9	- 2.5	- 1.8	- 2.0
	Gjesdal	1.3	0.5	- 2.5	- 3.9	- 1.1	- 1.3	- 3.1	- 3.0
	Bjerkreim	6.6	7.2	6.8	7.6	- 2.0	- 2.0	- 2.5	- 2.5
	1970-comm.								
	Stavanger	- 0.9	- 0.9	- 3.6	- 3.9	- 1.6	- 1.6	- 3.0	- 3.0
	Sandnes	1.1	1.0	- 4.1	- 3.0	0.0	- 0.1	0.8	0.7
	Forsand	1.5	0.9	12.6	10.5	- 3.7	- 3.7	- 2.1	- 2.2
	Gjesdal	1.5	1.5	0.5	1.2	- 1.2	- 1.3	- 2.8	- 2.8
	Bjerkreim	6.7	6.7	6.8	6.8	- 2.0	- 2.0	- 2.5	- 2.5
	Smallest common denominator								
10% limit	Stavanger Madla Hetland	- 0.9	- 1.1	- 3.6	- 3.8	- 1.6	- 1.8	- 3.0	- 3.0
10% limit	Sandnes Høle Høyland	1.1	1.6	- 4.1	- 3.8	- 0.0	0.7	0.8	0.8
10% limit	Forsand Gjesdal	1.5	1.5	3.8	3.4	- 1.7	- 1.6	- 2.5	- 2.6
10% limit	Bjerkreim	6.7	6.7	6.8	6.7	- 2.0	- 2.2	- 2.5	- 2.5
5% limit	Stavanger Sandnes Hetland Høyland Høle Madla	- 0.3	- 0.3	- 3.4	- 3.4	- 1.2	- 1.2	- 2.1	- 2.1
5% limit	Forsand Gjesdal	1.5	1.5	3.8	3.4	- 1.7	- 1.6	- 2.5	- 2.6
5% limit	Bjerkreim	6.7	6.4	6.8	6.7	- 2.0	- 2.2	- 2.5	- 2.5
No limit	Tot. area	- 0.2	-	- 2.9	-	- 1.3	-	- 2.2	-

than the differences measured as percentage point. This is to be expected, as the rates were calculated from absolute data values.

3. There are no definite differences between the two variables.

4. Disaggregation towards more units creates greater problems.

We shall now expand our data to all of Rogaland county, and only look at the first variable, number of people employed in industry.

In 1960 this area consisted of 54 communes, in 1970 the number was reduced to 26. To get an impression of what this really means when it comes to recalculation of data, we shall look at table 4.

Table 4: Types of changes between the communes of Rogaland in the time-period from 1960 to 1970

	Recalculations	
	From 1970-units to 1960-units	From 1960-units to 1970-units
Part of one commune transferred to another	44 (+1)	29 (-1)
Total aggregation of communes	0	16
	45	45

When we talk about errors inherent in our system, there are at least two complicating factors that have not yet been discussed. The first is the size of the coefficient used for recalculation, the second is the size of the actual data value transferred. These are intermingled and difficult to isolate from each other, because recalculation means multiplication of data value and coefficient. To illustrate:

1. If part of a commune is transferred to another one, the absolute figure added to the value of the receiving unit may introduce errors in the resulting data value according to the size-relation between the two figures. We can check this out by looking at the coefficient to be used for calculations the reverse way.
2. A coefficient for transference of 50 % is most liable to magnify potential errors, possibilities for errors are smaller at the extremes. It should be obvious that if our assumptions about homogeneity do not hold in the first place, the problem might be minimized by a small or a large coefficient for recalculation because then almost all or almost nothing is transferred between units.

Table 5 is a specification of table 4.

Table 5: Part transferences by size

Size of coefficient	From -70 to -60 units	From -60 to -70 units	Likely consequence
0-5/95-100%	13	13	Negligible errors
5-10/90-95%	7	4	Minor errors
10-25/75-90%	16	8	May introduce significant errors
25 - 75%	8	4	May introduce substantial errors

As we can see, in this case a relative larger part of the coefficients fall at the extremes if we calculate towards 70-units.

We have, in cases regarded relevant also done parallel tests for the common denominator units, but where we ignored all changes between communes that involved less than 10 % of the population of the commune that was reduced in size. If all changes should have been taken into consideration, we would get two distinctly different types of units. One half would be very big, while the reminder would be very small (especially islands) communes, left untouched by the upheaval of the sixties.

Test of relation between variables

To begin with, we shall look at correlations between the actual and the recalculated data values, as given in table 6.

Table 6: Correlation between actual and recalculated data

Correlations

	Actual level	Change percentage point	Change rate	
60-units 70-data	.925	.738	.702	N= 53
70-units 60-data	.999	.992	.994	N= 26

As we can see, recalculation of data (the "right" way towards fewer units) has minimal consequences for these correlations, even for the change-rates. This is the case even if some of the recalculations are quite complicated. Statistical analysis is not seriously effected by this kind of recalculations. If recalculation of data means disaggregation on a grand scale, however, measures of change tend to become inaccurate. If a correlation between two variables is .7, this indicates that only 50 % of the variance is common. After all, this leaves quite a lot of error, which certainly would effect whatever statistical analysis we undertake. This is a warning and deserve a closer look. It also strongly suggests that we should always try to recalculate towards fewer units.

If we correlate data for two or more variables recalculated from different points in time, it is a problem that we do not have the correct answer. However, we will assume that recalculation to fewer units gives almost correct answer. We shall now look at the correlation between number of people employed in industry and the election outcomes for the Social Democrats at the two elections 1961 and 1965. (All changes in commune borders in Rogaland between 1960 and 1970 took place between 1961 und 1965).

Table 7 presents correlations between the level-data of these variables on 1960- and 1970-units respectively.

Table 7: Correlations between employment in industry and election outcomes for Social Democrats

		Industry 60		Industry 70	
		Actual	Recalculated	Actual	Recalculated
60-units	S-61	.81		.78	.71
	S-65	.71		.72	.79
70-units	S-61	.83	.82	.77	
	S-65	.79	.78	.74	
Smallest common denominator	S-61		.83		.74
	S-65		.77		.70

The only conclusion we can draw from table 7 is that correlation between variables is only slightly effected by different kinds of recalculations. Maybe we can say that there is a small tendency that recalculation toward fewer units produces slightly higher coefficients, but the results are not unanimously pointing that way. The really important thing is, however, that correlations are only minimally effected.

In table 8 we have looked closer at what happens to the measures of change.

Table 8 clearly shows the problems we get with our measures of change if recalculation involves a general process of fission. If we assume that the real correlation is somewhere around .4 it is obvious that correlations as different as .25 and .62 certainly would lead us to different conclusions about the relation between the variables.

Table 8: Correlation between change in industrial employment and social democratic strength

	Change in industrial employment, percentage points 1960-1970		
		Actual	Recalculated
Change in social-democratic strength 1961-65, percentage points	1960-units	.25 [x]	.62
	1970-units	.40	.39
	Sm.C.D.	-	.44

[x] Actual means actual only for the industry variable.

We can try to explain why the two coefficients for the 1960-units differ so much to each side of the "correct" value. We know that the two variables really are positively correlated. If we recalculate data from 70-units to 60-units, we re-create many small rural communes which were added to a larger urban commune, and in this process we have to recalculate values of data from the value of the urban commune, and with a set of coefficients based on this larger unit.

The way our procedure works cannot quite re-create these rural communes with their former outlook, which would mean that we give them a positive, non-real, change-rate, they will be marked by the fact that they were merged with a commune quite different in outlook. If we use actual data, we have to remember that actual only means the industry variable. For the election outcomes we have no possibilities but to recalculate. This gives us one correct variable and one where we must expect a quite strong positive but non-real, change-rate. It is to be expected that if we recalculate two variables with the same procedure, we make them more similar and produce a higher correlation, than if we recalculate only one of them.

From table 7 we concluded that our recalculation-procedure worked well if the data specified the level of some unit. From table 8 we see that it is a lot more complicated if variables give measures of change. This kind of data should be treated with greater care.

Use of data across longer time-spans

A factor-analysis of a set of variables giving political, occupational and residential structure resulted in a 3-factor solution which takes out some 80 % of the variance of all the variables. Table 9 presents the communalities of the variables, percentage of the variance explained by 3 factors, and the eigenvalues of the three varimax-rotated factors, for seven different time-points.

It is clear that the overall pattern of the communalities is not effected by the recalculations of data, communalities are, however, a little lower if the number of units are high. The relative significance of the factors, measured as eigenvalues of the rotated factors are likewise only minimally altered, and in a systematic way.

Tables 10, 11 and 12 present the factor-loadings of all the variables on each of the factors. Again, the pattern is unchanged over different sets of units.

Our conclusion is that a factor-analysis run on this set of variables would lead to the same substantive conclusion whatever year we choose as the year for standardization of the units. Actual coefficients may be a little changed, but that would not lead the researcher to different conclusions about the underlying structure of the data.

Use of data for widely separate points in time

In 1919 there was a referendum in Norway and the Norwegians voted yes to prohibition. The result showed large regional variation and is usually explained as a "counter-cultural"-protest of the periphery. In 1972 there was another referendum, this time the question was Norway's entry into the European Common Market. This time the Norwegians voted no, and social scientists tend to start with the same explanation as mentioned above, it was a protest of the peripheries and the primary economic sector. We may then assume that there should be some positive correlation between the results. It is not easy to compare data from 1919 with data from 1972. We have done it on three different sets of units.

Figures 3-5 show three scattergrams with the two variables, of 1919-units (n=700), 1960-units (n=732), and 1972-units (n=444).

The scattergrams all show that there is a clear positive relation between the variables. Correlations are by and large not effected by the set on units we choose, the small differences may be explained just by differing number of units, an artifact of the calculation-formula.

CONCLUSIONS

The few and scattered empirical tests presented in this article do not warrant any strong conclusions, but we have tried to illustrate some of the difficult problems inherent in such a procedure. We have tried to illustrate what we regard as the most difficult situations (recalculation of communes

of very varied outlook across long time-spans, with many recalculations). We may conclude that the errors seldom are much greater than this study illustrates, and critical use of the system certainly can minimize these errors. On this background we may assume that recalculation of data should not seriously effect conclusions based on statistical analysis of all, or large subsets of the communes. It is more difficult to use long time series for just a few cases. Other types of variables may also be more difficult to handle. Our judgement at the moment is, however, that it is better use of resources at the present level to educate users than to work on the functioning of the system.

Table 9: Factoranalysis - Communality, percentage of variance for 3 factors, Eigenvalues

	1835 (n=362)	1865 (n=491)	1900 (n=594)	1930 (n=703)	1960 (n=732)	1970 (n=451)	1980 (n=454)
TURNOUT 73	.989	.981	.974	.965	.995	.980	.982
SOS.DEM. 73	.971	.954	.935	.927	.947	.984	.984
PRIMARY 60	.941	.897	.886	.867	.855	.939	.938
SOS.DEM. 61	.865	.839	.828	.810	.791	.874	.878
PRIMARY 70	.870	.830	.812	.800	.786	.852	.845
INDUSTRY 60	.751	.735	.726	.729	.727	.761	.766
RES.DENSITY 70	.837	.786	.762	.748	.727	.813	.802
RES.DENSITY 60	.766	.725	.705	.683	.663	.785	.773
TURNOUT 61	.767	.688	.656	.648	.619	.757	.763
NO-VOTERS 72	.680	.665	.649	.640	.615	.616	.616
INDUSTRY 70	.585	.565	.570	.572	.567	.589	.605
TURNOUT 72	.313	.335	.351	.338	.323	.329	.317
% of variance for three factors	82.4	80.3	79.4	78.7	77.8	81.9	81.8
Eigenvalues for rotated factors:							
FACTOR 1	6.14	5.86	5.80	5.69	5.56	6.01	6.01
FACTOR 2	1.74	1.66	1.64	1.61	1.62	1.75	1.76
FACTOR 3	1.45	1.48	1.41	1.42	1.44	1.52	1.50

Table 10: Factorloadings for 1. factor

	1835 (n=362)	1865 (n=491)	1900 (n=594)	1920 (n=703)	1960 (n=732)	1970 (n=451)	1980 (n=454)
RES.DENSITY 70	.91	.88	.87	.86	.84	.90	.89
RES.DENSITY 60	.87	.84	.83	.82	.81	.88	.87
INDUSTRY 60	.80	.80	.80	.80	.80	.81	.81
INDUSTRY 70	.69	.69	.69	.69	.69	.71	.73
TURNOUT 61	.37	.33	.34	.32	.30	.31	.32
SOS.DEM 61	.21	.22	.23	.23	.22	.21	.21
TURNOUT 73	.15	.13	.15	.14	.13	.14	.15
SOS.DEM 73	.08	.10	.11	.11	.11	.12	.12
TURNOUT 72	.09	.11	.09	.10	.09	.11	.10
NO-VOTERS 72	-.68	-.66	-.65	-.62	-.61	-.63	-.61
PRIMARY 70	-.90	-.88	-.87	-.86	-.85	-.89	-.88
PRIMARY 60	-.96	-.93	-.93	-.92	-.91	-.95	-.95

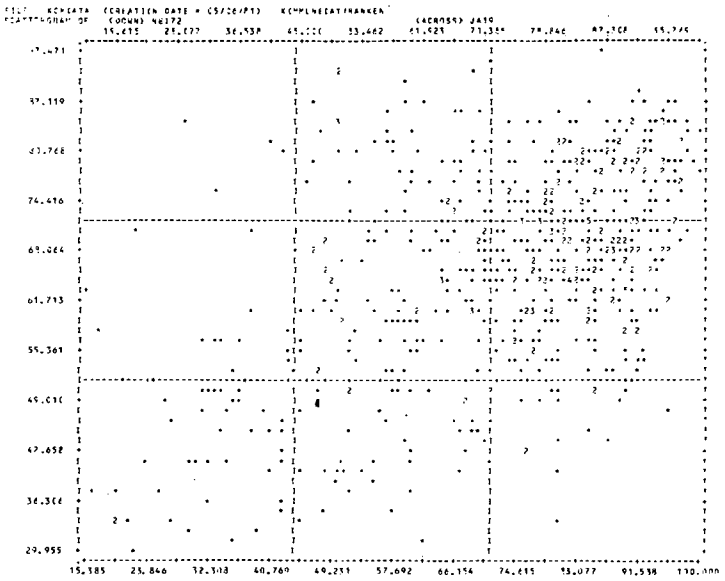
Table 11: Factorloadings for 2. factor

	1835 (n=362)	1865 (n=491)	1900 (n=593)	1920 (n=703)	1960 (n=732)	1970 (n=451)	1980 (n=454)
TURNOUT 73	.97	.97	.97	.97	.98	.98	.98
TURNOUT 61	.75	.72	.69	.70	.69	.78	.78
TURNOUT 72	.51	.52	.53	.52	.51	.51	.50
INDUSTRY 70	.32	.30	.29	.31	.31	.28	.28
INDUSTRY 60	.27	.26	.23	.24	.23	.28	.29
SOS.DEM. 73	.12	.12	.10	.12	.11	.09	.09
RES.DENSITY 70	.08	.06	.08	.09	.08	.05	.05
RES.DENSITY 61	.04	.04	.04	.03	.02	.04	.03
SOS.DEM. 61	-.05	-.06	-.07	-.07	-.07	-.06	-.07
PRIMARY 60	-.12	-.13	-.11	-.11	-.10	-.16	-.16
PRIMARY 70	-.22	-.21	-.21	-.23	-.23	-.23	-.23
NO-VOTERS 72	-.37	-.38	-.36	-.39	-.38	-.38	-.39

Table 12: Factorloadings for 3. factor

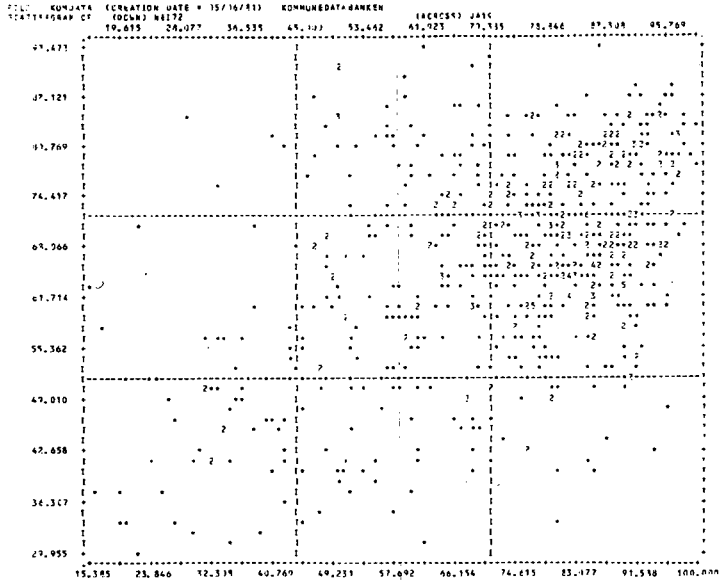
	1835 (n=362)	1865 (n=491)	1900 (n=594)	1920 (n=703)	1960 (n=732)	1970 (n=451)	1980 (n=454)
SOS.DEM. 73	.97	.96	.95	.95	.96	.98	.98
SOS.DEM. 61	.90	.89	.88	.87	.86	.91	.91
TURNOUT 61	.27	.25	.25	.25	.24	.24	.23
INDUSTRY 60	.18	.17	.18	.17	.17	.15	.15
RES.DENSITY 60	.07	.10	.11	.11	.11	.10	.12
TURNOUT 73	.16	.12	.10	.11	.10	.08	.08
RES.DENSITY 70	.05	.07	.07	.09	.09	.09	.10
INDUSTRY 70	.09	.05	.06	.04	.04	.04	.04
PRIMARY 60	-.04	-.09	-.10	-.10	-.10	-.08	-.08
PRIMARY 70	-.11	-.11	-.12	-.12	-.12	-.12	-.13
TURNOUT 72	-.20	-.23	-.24	-.24	-.24	-.23	-.23
NO-VOTERS 72	-.29	-.29	-.30	-.33	-.32	-.28	-.29

Figure 3: Scattergram for no-voters 1972 with yes-votes 1919 on 1919-units (n=700)



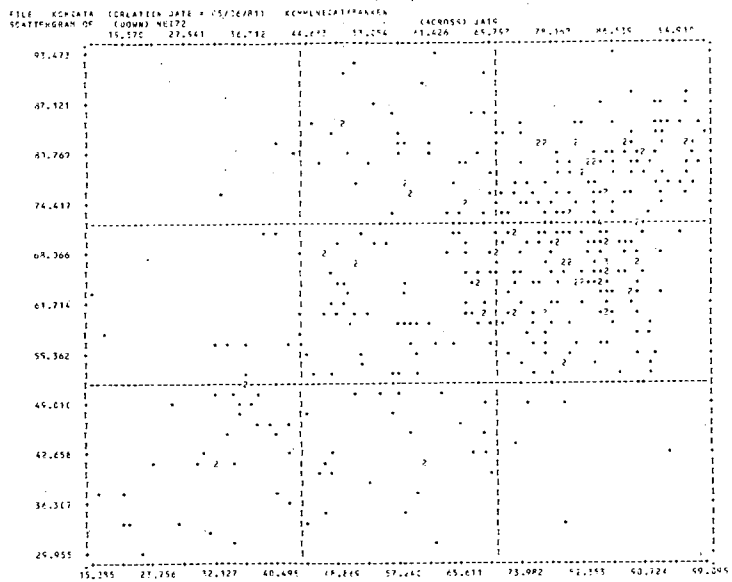
R= .48 b= .33

Figure 4: Scattergram for no-voters 1972 with yes-votes 1919 on 1960-units (n=732)



R= .46 b= .32

Figure 5: Scattergram for no-voters 1972 with yes-votes 1919 on 1972-units (n=444)



R= .59 b= .36